Emergent Semantics from Folksonomies: A Quantitative Study

Lei Zhang, Xian Wu, and Yong Yu

APEX Data and Knowledge Management Lab,
Department of Computer Science and Engineering,
Shanghai JiaoTong University, Shanghai, 200030, China
{zhanglei, wuxian, yyu}@apex.sjtu.edu.cn

Abstract. Defining and using ontology to annotate web resources with semantic markups is generally perceived as the primary way to implement the vision of the Semantic Web. The ontology provides a shared and machine understandable semantics for web resources that agents and applications can utilize. This top-down approach (in the sense that an ontology is defined first on top of existing web resources and then used later to markup them), however, has a high barrier to entry and is difficult to scale up. In this paper, we investigate using a bottom-up approach for semantically annotating web resources as supported by the now widely popular social bookmarks services on the web where users can annotate and categorize web resources using "tags" freely choosen by the user without any pre-existing global semantic model. This kind of informal social categories is coined as "folksonomies". We show how global semantics can be statistically inferred from the folksonomies to semantically annotate the web resources. The global semantic model also disambiguate the tags and group synonymous tags together. Finally, we show that there indeed are hierarchical relations among the emerged concepts in the folksonomy and it is plausible to further identify them if we use more advanced probabilistic models.

1 Introduction

Semantic Web is a vision that web resources are made not only for humans to read but also for machines to understand and automatically process [1]. This requires that web resources be annotated with machine understandable metadata. Currently, the primary approach to achieve this is to firstly define an ontology and then use the ontology to add semantic markups for web resources. These semantic markups are written in standard languages such as RDF [2] and OWL [3] and the semantics is provided by the ontology that is shared among different web agents and applications. We refer to this approach as the top-down approach because an global semantic model (i.e., the ontology) is defined and imposed on top of web resources before we actually use the semantic model to annotate these resources.

The top-down approach has several drawbacks. Firstly, establishing an ontology as a semantic backbone for a large number of distributed web resources is

S. Spaccapietra et al. (Eds.): Journal on Data Semantics VI, LNCS 4090, pp. 168–186, 2006.

not easy. Different people/applications may have different views on what exists in these web resources and this leads to the difficulty of the establishment of and commitment to a common ontology. Even if the consensus of a common ontology can be achieved, it may not be able to catch the fast pace of change of the targeted web resources. A lot of work has been done on developing ontology engineering tools to help people create ontologies, such as Protégé [4], OilEd [5], WebODE [6], ORIENT [7] and SWOOP [8]. While these tools facilitate the actual construction of ontologies, they generally do not help much in forming the required consensus for ontolgy building in a distributed environment. Using these tools also requires some level of expertise in ontology engineering or knowledge engineering, which put a high barrier to entry for the mass developers and users. Studies on ontology evolution, such as [9,10], focus on how changes of ontologies are tracked [11,12], versioned [13,14] and managed [15] but does not provide mechanisms to automatically and actively change the ontology according to the changes of web resources it intends to cover. Secondly, even if we have successfully built an ontology, using it to make semantic annotations in an automatic and scalable manner is still a challenging task. Usually, the semantic annotations are made manually [16,17] or semi-automatically [18,19,20,21]. Although this helps create high quality semantic annotations, it is hard to scale up. Till now, only very little work has been done on large-scale fully automatic semantic annotations of web reseources [22,23,24].

The above shortcomings of the top-down approach have actually already been identified in the "emergent semantics" research [25,26] in which semantics is treated as an agreement that is achieved in a bottom-up and incremental manner without relying on pre-existing global semantic models. In this paper, we investigate whether and how the semantic annotation problem can be attacked in this bottom-up emergent semantics way. Our work is enabled and supported by the now widely popular social bookmarks services on the web, like Delicious[1], Furl[2] and Yahoo My Web 2.0[3]. These services allow web users to annotate and categorize web resources using "tags" that are freely chosen by the user without any "a-priori" dictionary, taxonomy, or ontology to conform to. Thus, the tags can be any strings that the user deems appropriate for the web resource. The name "folksonomy" has been coined for this kind of informal social categorization of web resources. In our view, this is also a massive bottom-up annotation of web resources that directly complements the traditional top-down approach of semantic annotation. If emergent semantics can be derived from these free-style bottom-up annotations, it will remedy the headache of top-down approach to semantic annotations. It removes the high barrier to entry because web users can annotate web resources easily and freely without using or even knowing taxonomies or ontologies. It directly reflects the dynamics of the vocabularies of the users and thus evolves with the users. It also decomposes the burden of annotating the entire web to the annotating of interested web resources by each individual web users.

[1] `http://del.icio.us`
[2] `http://www.furl.net`
[3] `http://myweb2.search.yahoo.com`

Apparently, without a shared taxonomy or ontology, the folksonomy suffers the usual problem of ambiguity of semantics. The same tag may mean different things for different people and two seemingly different tags may bear the same meaning. Without a clear semantics, these bottom-up annotations won't be much useful for web agents and applications on the Semantic Web. In this paper, we propose to use a probabilistic generative model to model the user's annotation behavior and to automatically derive the emergent semantics of the tags. Synonymous tags are grouped together and highly ambiguous tags are identified and separated. Finally, we show that we can use more advanced probabilistic models to discover the hierarchical relations among the emerged concepts in the folksonomy.

2 Folksonomy

The idea of a bottom-up approach to the semantic annotation is enlightened and enabled by the now widely popular social bookmarks services on the web. These services provide easy-to-use user interfaces for web users to annotate and categorize web resources, and furthermore, enable them to share the annotations and categories on the web. For example, the Delicious (`http://del.icio.us`) service allows you to easily add sites you like to your personal collection of links, to categorize those sites with keywords, and to share your collection not only between your own browsers and machines, but also with others. There are many bookmarks manager tools available [27,28]. What's special about the social bookmarks services like Delicious is their use of keywords called "tags" as a fundamental construct for users to annotate and categorize web resources. These tags are freely chosen by the user without a pre-defined taxonomy or ontology. Some example tags are "blog", "mp3", "photography", "todo" etc. The tags page[4] of the Delicious web site lists most popular tags among the users and their relative frequency of use. These user-created categories using unlimited tags and vocabularies was coined a name "folksonomy" by Thomas Vander Wal in a discussion on information architecture[5]. The name is a combination of "folk" and "taxonomy".

As pointed out in [29], folksonomy is a kind of user creation of metadata which is very different from the professional creation of metadata (e.g. created by librarians) and author creation of metadata (e.g. created by a web page author). Without a tight control on the tags to use and some expertise in taxonomy building, the system soon runs into problems caused by ambiguity and synonymy. [29] cited some examples of ambiguous tags and synonymous tags in Delicious. For example, the tag "ANT" is used by many users to annotate web resources about Apache Ant, a building tool for Java. One user, however, uses it to tag web resources about "Actor Network Theory". Synonymous tags, like "mac" and "macintosh", "blog" and "weblog" are also widely used. What's more important about folksonomies is that the tags are all in a flat namespace without hierarchy or any parent-child relationships.

[4] `http://del.icio.us/tag/,accessedatNovember2005.`

[5] `http://atomiq.org/archives/2004/08/folksonomy_social_classification.html,accessed at November 2005.`

Despite of the seemingly chaos of unrestricted use of tags, social bookmarks services still attract a lot of web users and provide a viable and effective mechanism for them to organize web resources. [29] contributes the success to the following reasons.

- Low barriers to entry
- Feedback and asymmetric communications
- Individual and community aspects

Unlike the professional creation of metadata or the top-down approach of the semantic annotation, folksonomy does not need sophisticated knowledge about taxonomy or ontology to do annotation and categorization. This significantly lowers the barrier to entry. In addition, because these annotations are shared among all users in a social bookmark service, there is an immediate feedback when a user tags a web resource. The user can immediately see other web resources tagged by other users using the same tag. These web resources may not be what the user expected. In that case, the user can adapt to the group norm, keep the tag in a bid to influence the group norm, or both [30]. Thus, the users of folksonomy are negotiating the meaning of the terms in an implicit asymmetric communication. This local negotiation, from the emergent semantics perspective, is the basis that leads to the incremental establishment of a common global semantic model. [31] made a good analogy with the "desire lines". Desire lines are the foot-worn paths that sometimes appear in a landscape over time. The emergent semantics is like the desire lines. It emerges from the actual use of the tags and web resources and directly reflects the user's vocabulary and can be used back immediately to serve the users that created them. In the following of the paper, we quantitatively analyze the folksonomy and show that emergent semantics indeed can be inferred statistically from it.

3 The Data of Social Bookmarks

Social bookmark services can provide many functionalities for end users. Different services may have different functions. Some allow users to give a short description of each bookmark. Some allow users to rate each bookmark for its quality. These different functions acquire different kind of data from end users for web bookmarks. In this paper, we focus on the most important data that are common to most social bookmarks. The core function of a social bookmark service is to let users bookmark URLs and assign tags to URLs. Tags are words or phrases that are freely chosen by users. This core function is common to most social bookmark services. Hence, in this paper, we focus on this core function and the data asscoicated with it, namely the user, the URL and the tag.

3.1 Co-occurrence Data Model

We abstract the data in social bookmarks services as a set of quadruples

$$(user, URL, tag, time)$$

which means that a user tags a URL with a specific tag at a specific time. In this paper, we focus more on what URL gets what tags and ignore the user and time information in the quadruple. What interests us is thus the co-occurrence of tags and URLs. Let's denote the set $X = \{x_1, x_2, \ldots, x_N\}$ and $Y = \{y_1, y_2, \ldots, y_M\}$ to be the set of URLs and the set of tags in the collected folksonomy data respectively. Each quadruple then translates to a co-occurence of a URL and a tag. The set of quadruples then translates to the co-occurrence set $S = \{(x_{i(r)}, y_{j(r)}, r) : 1 \leq r \leq L\}$. L is the total number of co-occurrences/pairs in S. $x_{i(r)}$ corresponds to the URL in X which appears in the r^{th} pair. $y_{j(r)}$ corresponds to the tag in Y which appears in the r^{th} pair. $n_{ij} = |\{(x_i, y_j, r) \in S\}|$ measures the frequency of co-occurrence of URL x_i and tag y_j.

We have collected a sample of Delicious data by crawling its web site during March 2005. The data set consists of 2,879,614 taggings made by 10,109 different users on 690,482 different URLs with 126,304 different tags. The co-occurrence data can be easily computed from the raw dataset. The following paper will use the dataset for experiments.

3.2 Social Aspects: The Power Law

The biggest difference between a set of personal bookmarks and a social bookmarks service is the implicit social interactions enabled by the latter. Typically, users of a socical bookmarks service can see other users' public bookmarks and tags. For a given tag (or a set of tags), users can see what URLs other users have tagged using the same tag(s). This function is very valuable for the users because it enables them to discover potentially high-quality web resources collected by other users of the same topic. When the user bookmarks a URL, tags used by other users for the same URL can also be seen. This may influence the user on what tag(s) to use for bookmarking the URL. Because these functions of the social bookmarks service are both very valuable and interesting, users of the service frequently use these functions, which is actually implicit social interactions. As we have analyzed in section 2, through these implicit social interactions, users are negotiating the meanings and uses of tags on URLs. These local negotiations, from the emergent semantics perspective, enable the incremental establishment of a common global semantic model.

When a lot of users are involved in the implicit social interactions, interesting phenomenons emerge. If an URL is bookmarked by many users, it has more chance to be seen by other users. The more chance to be seen by other users, the more the URL may be bookmarked. This positive loop will lead to an exponential growth of the number of the times an URL being bookmarked. Tags have the similar situation. If a common tag is used by many people for tagging many URLs, it has more chance to be seen by other users. The more chance for the tag to be seen, the more it may be used by users to tag more URLs. This is also a self-rewarding positive loop. Similar situations also occur on the web. If a web page is linked by many other pages, it has more chance to be seen by users. The more chance to be seen by users, the more chance it may be linked by more web pages. On the web, this phenomenon is reflected in the distribution

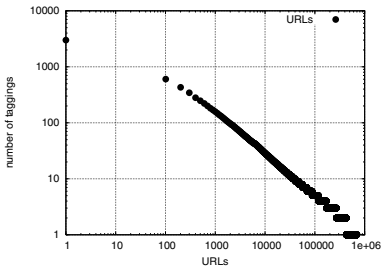

Fig. 1. The distribution of the taggins of URLs

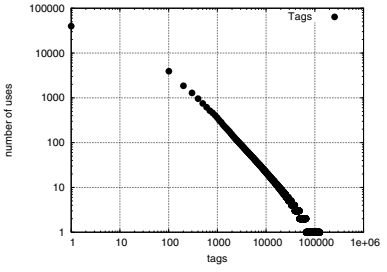

Fig. 2. The distribution of the uses of tags

of the in-bound links of web pages. Only very few pages have very large amount of in-bound links and most web pages only have a few in-bound links. Study showes that the growth of the web follows the Power Law [32], meaning that the probability of attaining a certain size x is proportional to $1/x$ to a power β, where β is greater than or equal to 1.

We expect that the social bookmarks data also has the Power Law distribution. To verify this, using the Delicious data set we collected, we computed the distribution of the number of taggings of URLs and the number of uses of tags. More precisely, for every URL $x_i \in X$, we computed the number of taggings users have made on it: $n_{x_i} = |\{(x_i, y, r) \in S\}|$. For every tag $y_j \in Y$, we computed the number of uses of the tag by all the users: $n_{y_j} = |\{(x, y_j, r) \in S\}|$. Fig.1 and Fig.2 show the results of the two computed distributions respectively[6]. Since both the axes of the figures are in log-scale, the figures clearly show Power Law distributions. This reflects the implicit social interactions inherent in the social bookmarks service.

4 Deriving Emergent Semantics

4.1 Probabilistic Generative Model

The co-occurrences of URLs and tags is not a random phenomenon. It reflects the underlying semantics that users has assigned to these URLs and tags. We propose use the following probabilistic generative model to model the user's behavior in assigning a tag to a URL. The model assumes the exist of a set of concepts $C = \{c_1, c_2, \ldots, c_K\}$.

1. User randomly encounters a URL x_i on the web with probability p_i .
2. The URL makes the user thinking of a concept c_α with probability $p_{\alpha|i}$.
3. The concept c_α trigers the user to use tag y_j with probability $p_{j|\alpha}$.

Here, both $p_{\alpha|i}$ and $p_{j|\alpha}$ are conditional probabilities. $p_{\alpha|i}$ is the probability of thinking of concept c_α given the URL x_i. $p_{j|\alpha}$ is the probability of using tag y_j given the concept c_α. This probabilistic generative model can be visually

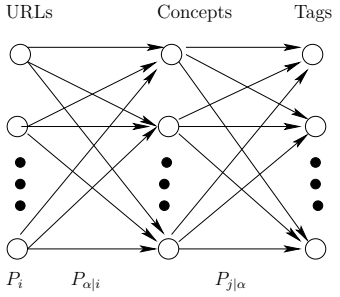

Fig. 3. Probabilistic generative model

depicted as Fig.3 The model makes a simplified independence assumption that once a concept is thought by a user, the tag to use is only determined by the concept and is indenpendent of the URL that trigers the concept. Note that the set of concepts is actually the underlying semantics that controls the co-occurrences of URLs and tags. The problem is then how to get the set of concepts and their probability relations with the tags and URLs. Directly estimates the probabilities is very difficult. The set of URLs X is potentially very large because of the overwhelming size of the web . The set of tags Y could also be very large because folksonomy has no control on the use of tags. Any string could be a tag. Thus, the frequency of a pair (x_i, y_j) may be very very low and this creates the data sparseness problem for model parameter estimation. However, the introduce of the concept set C remedies the problem. Hofmann and Puzicha [33] proposed a EM algorithm for estimating the parameters. The above model corresponds to the asymmetric SMM model for co-occurrence data in [33].

Hofmann and Puzicha showed [33] that the asymmetric SMM model is equivalent to its symmetric version: the SMM model which is easier to compute and can handle larger dataset. In our case, the SMM model can be explained as the following probabilistic model:

1. The user randomly thought of a concept c_α with probability π_α .
2. A URL x_i is selected by the user for the concept c_α with probability $p_{i|\alpha}$.
3. A tag y_j is selected by the user for the concept c_α with probability $q_{j|\alpha}$

The x_i and y_j are conditionally independent given the concept c_α and the joint probability distribution of the SMM is a mixture of separable component distributions (hence the name, Separable Mixture Model) which can be parameterized by

$$p_{ij} = P(x_i, y_j) = \sum_{\alpha=1}^{K} \pi_\alpha P(x_i, y_j | c_\alpha) = \sum_{\alpha=1}^{K} \pi_\alpha p_{i|\alpha} q_{j|\alpha}$$

Following the EM approach, to optimally fit the SMM model to the observation of co-occurrences set S, and estimate the parameters, the log-likelihood of each pair co-occurrences probability $\left(p_{ij}^{n_{ij}}\right)$ for all pairs

[6] In order to reduce the size of the figures' EPS file, only 1/100 data points are drawn.

$$L = \sum_{i=1}^{N} \sum_{j=1}^{M} n_{ij} log \left(\sum_{\alpha=1}^{K} \pi_\alpha p_{i|\alpha} q_{j|\alpha} \right)$$

should be maximized. As a standard method for EM algorithm for mixture models, a hidden variable $R_{r\alpha}$ is introduced which denotes the probability that the observation $\left(x_{i(r)}, y_{j(r)}, r\right)$ is generated from the concept c_α. The EM method leads to the

E-Step

$$\langle R_{r\alpha} \rangle^{(t+1)} = \frac{\hat{\pi}_\alpha^{(t)} \hat{p}_{i(r)|\alpha}^{(t)} \hat{q}_{j(r)|\alpha}^{(t)}}{\sum_{v=1}^{K} \hat{\pi}_v^{(t)} \hat{p}_{i(r)|v}^{(t)} \hat{q}_{j(r)|v}^{(t)}}$$

M-Step

$$\hat{\pi}_\alpha^{(t)} = \frac{1}{L} \sum_{r=1}^{L} \langle R_{r\alpha} \rangle^{(t)}$$

$$\hat{p}_{i|\alpha}^{(t)} = \frac{1}{L\hat{\pi}_\alpha^{(t)}} \sum_{r:i(r)=i}^{L} \langle R_{r\alpha} \rangle^{(t)}$$

$$\hat{q}_{j|\alpha}^{(t)} = \frac{1}{L\hat{\pi}_\alpha^{(t)}} \sum_{r:j(r)=j}^{L} \langle R_{r\alpha} \rangle^{(t)}$$

Iterating the E-Step and M-Step, the parameters converge to a maximum of the likelihood. Our collected raw Delicious data is very large for the EM algorithm. We made a random sample of the collected Delicious data. The sample has 17,707 URLs, 7,238 tags and 300,869 co-occurrences in total. We set the number of concepts to 50 and run through the EM algorithm of the SMM model. After computation, the parameter $q_{j|\alpha}$ gives the conditional distribution of tags over the 50 concepts. We selected the top 10 concepts and for each concept the first five tags that have the highest $q_{j|\alpha}$ value. The result is shown in Table 1. We can see that tags that have the same semantics are effectively grouped together in one concept. The concepts thus can be seen as a "classes" in an ontolgy or "synsets" in WordNet [34].

4.2 Emergent Semantics

Using the results obtained by the probability generative model, we can derive and represent the emergent semantics of URLs and tags. For a given URL, its semantics should be represented by the concepts the URL is related to. Let's use $p_{\alpha|i}$ to denote the conditional probability that a concept c_α is thought of by the user given an URL x_i. For a given URL x_i, the $p_{\alpha|i}$ values for all concepts c_α actually represents a discrete probability distribution on all the concepts. This distribution describes in detail the concepts that the URL relates to and the strength of the relatedness. We thus use this distribution as the representation

Table 1. Concepts and Tags

Concept	Top 5 tags in the concept
1	technology Google Search Internet future
2	Php PHP webdev mysql code
3	programming development Programming cs toread
4	del.icio.us delicioius bookmarks tags folksonomy
5	humor fun humour Funny ukquake
6	software windows tools Software freeware
7	books book library literature copyright
8	bittorrent p2p torrents BitTorrent P2P
9	comics comic humor webcomic Comics
10	security wordpress hack wifi Security

of the semantics of the URL. Since it is a discrete distribution, we can represent it as a vector. The semantics of a URL x_i is thus represented as

$$\overrightarrow{semantics(x_i)} = \left\langle\, p_{\alpha|i} \mid \alpha = 1, 2, \ldots, K \right\rangle$$

where $p_{\alpha|i}$ can be computed as follows using Bayesian theorem:

$$p_{\alpha|i} = \frac{p_{i|\alpha}\pi_\alpha}{p(x_i)} = \frac{p_{i|\alpha}\pi_\alpha}{\sum_{\alpha=1}^{K} p_{i|\alpha}\pi_\alpha}$$

π_α and $p_{\alpha|i}$ have been obtained via the EM algorithm in the probabilistic generative model. Therefore, the representation of the semantics of a URL can be computed.

Using the previous experiment data, we calculated the semantic representations of all URLs in the data. Fig.4 to Fig.7 show the concept distributions of four URLs. URL-1 is a special URL used by the Delicious service for replacing all ill-formated URLs users have bookmarked. Since there is a great variety of different ill-formated URLs and their tags, this special URL has no prominent concepts associated with it. This is also reflected in its concept distribution in Fig.4 where the URL is related to almost all concepts in very low strength (< 0.07). In contrast, the other three URLs all have prominent concepts. URL-2 to URL-4 are `http://www.yahoo.com`, `http://jakarta.apache.org` and `http://www.filelist.org` respectively. Their concept distributions all have spikes that have strong relatedness to the URL (> 0.6).

Similar to the representation of the semantics of an URL, we can define the representation of the semantics of a tag y_j as:

$$\overrightarrow{semantics(y_j)} = \left\langle\, q_{\alpha|j} \mid \alpha = 1, 2, \ldots, K \right\rangle$$

where $q_{\alpha|j}$ is the conditional probability that a concept c_α is generated given the tag y_j. This probability can be computed as

$$q_{\alpha|j} = \frac{q_{j|\alpha}\pi_\alpha}{p(y_j)} = \frac{q_{j|\alpha}\pi_\alpha}{\sum_{\alpha=1}^{K} q_{j|\alpha}\pi_\alpha}$$

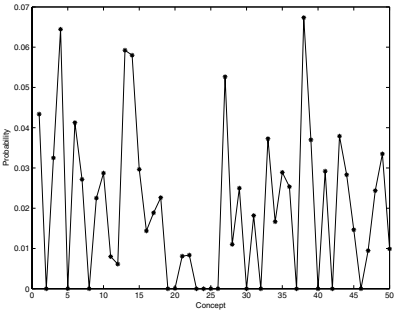

Fig. 4. Distributions of URL-1

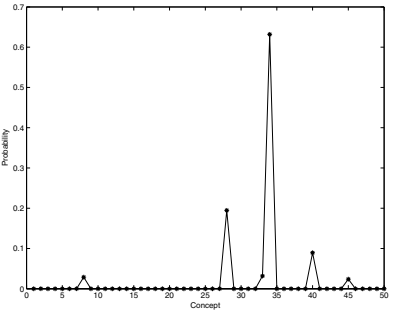

Fig. 5. Distributions of URL-2

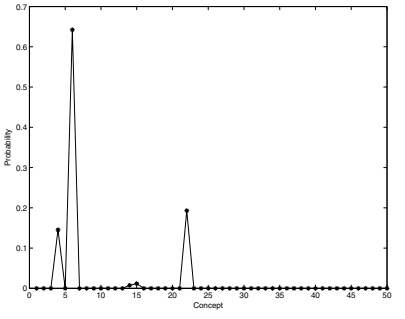

Fig. 6. Distributions of URL-3

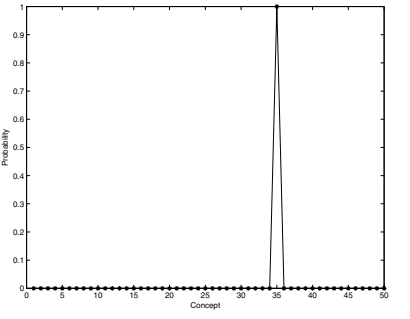

Fig. 7. Distributions of URL-4

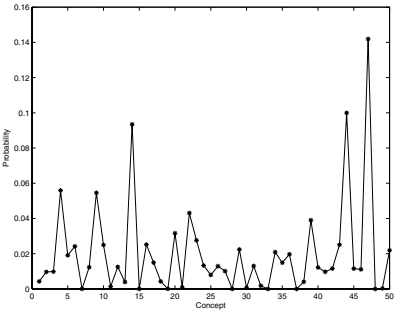

Fig. 8. Distribution of "todo"

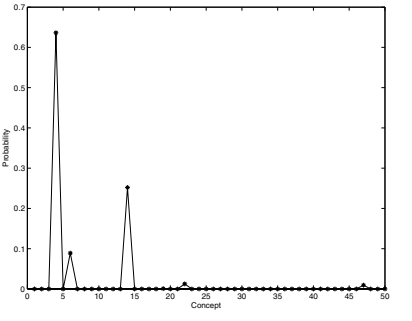

Fig. 9. Distribution of "xp"

where both $q_{j|\alpha}$ and π_α have been obtained in the probabilistic generative model. Therefore we can also compute the semantic representation of a tag. Using the previous experiment data, we calculated four tags' semantic representations. The result is shown in Fig.8 to Fig.11.

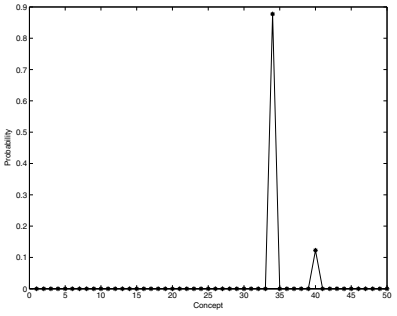

Fig. 10. Distribution of "google"

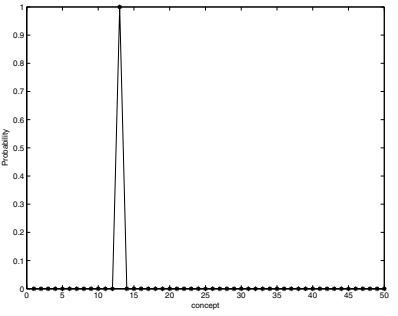

Fig. 11. Distribution of "cooking"

The tags "todo" and "cooking" are two extreme cases. Because what to do next is vastly different for different people, the tag "todo" is used to mark a lot of different URLs for different meanings of what to do next. This makes the "todo" tag very ambiguous. This is reflected in its concept distribution in Fig.8. On the contrary, the tag "cooking" is used very unambiguously in our experiment data set. Thus, its concept distribution as shown in Fig.11 only has one very big spike. The other two tags, "xp" and "google", are between the two extreme cases. For tag "xp", it is mainly used for the meanings of "windows xp" or "extrem programming". Likewise, the tag "google" is mostly used together with "search" or "gmail" for the meaning of internet search or google gmail. The two tags' concept distributions (as in Fig.9 and Fig.10) therefore has two or more spikes.

The above examples have shown the clear difference between ambiguous tags/URLs and unambiguous ones. Their concept distributions (or equivalently, their semantic representations) have very different characteristics. The concept distributions of ambiguous tags/URLs are more evenly distributed while those of unambiguous ones usually have very prominent spikes. This leads us to the idea of quantitatively measure the ambiguousness of a tag/URL using the entropy of its concept distribution. The ambiguousness of a tag/URL thus can be seen as a function of its semantic representation. More precisely, we define the ambiguity of a URL x_i and/or a tag y_j as follows:

$$ambiguity(x_i) = -\sum_{\alpha=1}^{K} p_{\alpha|i} \log p_{\alpha|i}$$

$$ambiguity(y_j) = -\sum_{\alpha=1}^{K} q_{\alpha|j} \log q_{\alpha|j}$$

where $p_{\alpha|i}$ and $q_{\alpha|j}$ are exactly the dimension value within the vectors $\overrightarrow{semantics(x_i)}$ and $\overrightarrow{semantics(y_j)}$. Using this definition, we calculated the ambiguity of all tags in the experiment data set and the result is shown in Table 2. The table shows the top 10 tags with the largest and smallest ambiguity

Table 2. Tags and their entropy

NO.	Tags	Ambiguity	Tags	Ambiguity
1	todo	3.24	cooking	0
2	viapopular	3.19	webmail	0
3	.imported	3.18	Deutsch	0
4	temp	3.08	netlabel	0
5	linklog	3.07	OWL	0
6	new	3.05	ttf	0
7	resources	3.04	vegetarian	0
8	from/furl	3.03	Sudan	0
9	resource	3.02	dictionary	0
10	[en]	3.00	rgb	0

values in column two and four respectively. In addition to "todo", we noticed that the tags "viapopular", ".imported", and "from/furl" are also very ambiguous. These tags are used to mark URLs imported from other bookmarks, which basically does not restrict the meaning of the tags to any specific concept. General tags, like "new", "resource" and "[en]" also appears ambiguous because they are too general to mean any particular concept. Note that the ambiguous word "OWL" appears very unambiguously in the list because the Delicious community is mostly concerned with IT technology. Thus, "OWL" in Delicious does not mean the bird of night but the web ontology language OWL. Hence, this tag appears very unambiguously.

In this subsection, we have defined the representation of the semantics of a tag/URL as a concept vector that corresponds to a discrete concept distribution of the tag/URL. We'd like to emphasize that this semantic representation is very different from the ontology-based top-down approach to semantic annotation. In the top-down approach, ontology is built beforehand whereas in our bottom-up approach the set of concepts is dynamically determined from the data set via a probabilistic model. Traditional semantic annotation is basically a binary judgement. An object is either an instance of a concept or not. However, in our model, the semantics of a tag/URL is not a binary classification but a discrete probability distribution over all the concepts. Compared with binary classification, this representation can better accommodate the inaccuracy, fuzziness and ambiguity of semantics. We have shown how ambiguity can be computed from the semantic representation. This semantic representation is also a computational result of the data set, that is, it is emerged rather than assigned. In the above example, the "OWL" tag is currently unambiguous in the data. When users are going to use "OWL" to mean more and more about other things, e.g. the night bird, its computed semantic representation from the data will change accordingly to accommodate new meanings. This is the real power of the emergent semantics. It dynamically reflects the current state of the system and evolves with it.

Compared to the top-down approach of semantics, what we currently lacking is a hierarchy structure of the emerged concepts. Well-organized hierarchy structures of concepts is a strong point of the top-down approach. In the

following subsection, using a more refined probabilistical model, we show that there indeed **are** hierarchical relations among the emerged concepts.

4.3 Hierarchical Concept Relations

In order to find a hierarchy of all the concepts hidden in the tags, we utilized the HACM model in [33]. HACM is a hierarchy clustering model. Fig.12 from [33] shows the schema for data generation in HACM model. The rectangle at the

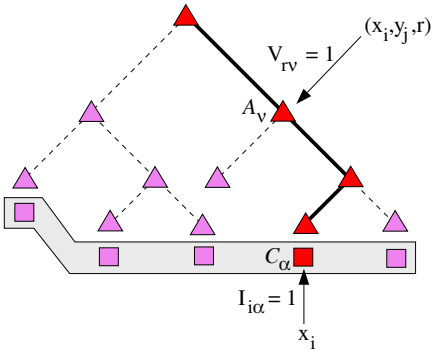

Fig. 12. Schema for data generation in HACM

bottom represents the concepts like in the SMM model. Triangle nodes denote inner nodes of a hierarchy. In the folksonomy scenario, the users' tagging behavior can be explained using the HACM model as follows:

1. The user encounters some URL x_i with probability p_i.
2. The URL makes the user think of one concept c_α in the bottom of the hierarchy. A hidden binary variable $I_{i\alpha}$ is used to denote which concept is chosen for the URL.
3. The user selects a generalization level v for the concept. This generalization level determines an inner node in the path from the concept c_α at the bottom to the root node at the top. A hidden binary variable V_{rv} is introduced to encode the resolution level $\mathcal{A}_v$ for the r^{th} co-occurrence observation.
4. A tag y_j is chosen given the inner node $\mathcal{A}_v$ with probability $q_{j|\alpha}$.

Note that the major difference with previous generative models is that the user has to select a generalization level before generate the tag from the assigned concept. Here, we omit the mathematical details and the EM algorithm of the HACM model. Interested readers are referred to [33] for further reading.

We experimented using HACM model to automatically generate hierarchy structures from the Delicious data we collected. In the experiment, we assumed a complete binary tree structure. We are well aware that this is a radical simplification and bold assumption because concept hierarchies need not to be so. The concept hierarchy can even not be a tree but be a lattice. The purpose

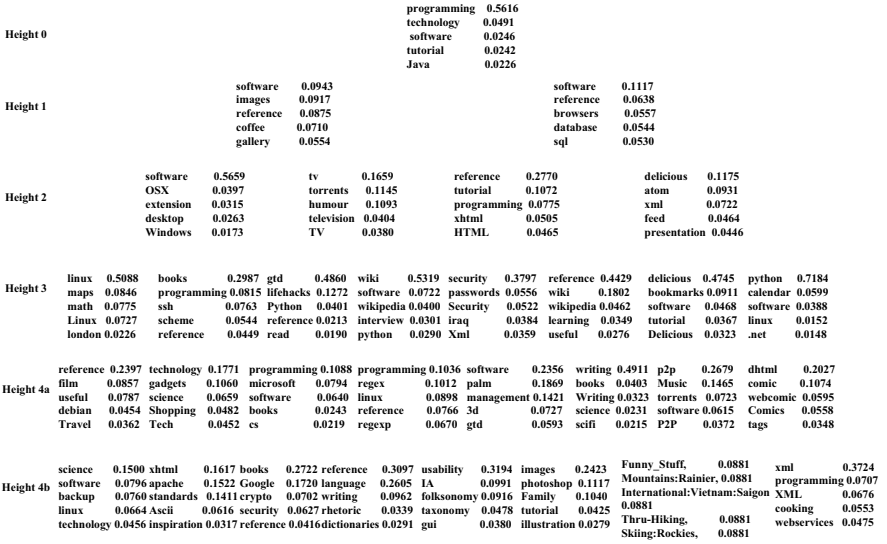

Fig. 13. Automatically generated taxonomy

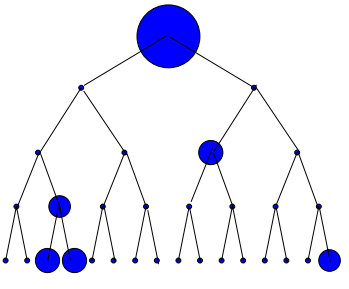

Fig. 14. The distribution of the tag "programming"

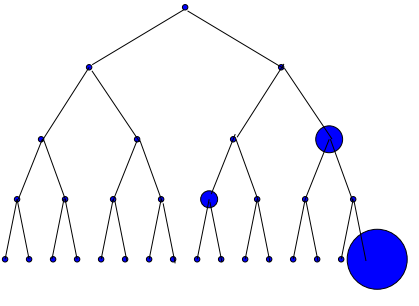

Fig. 15. The distribution of the tag "xml"

of this experiment, thus, is not to prove the correctness or robustness of the method to derive concept hierarchies but to quantitatively study whether there are narrower-broader relationships among the emerged concepts. We randomly sampled the raw data to get a small test data with 1642 URLs and 1121 tags co-occurred for 37,124 times to speed up this experiment. Fig.13 shows one of the results of the experiments. The depth of the taxonomy is set to 5. The last two rows of Fig.13 are actually at the same height 4 but are wrapped to fit the page size. The numbers at the right size of the tags is the probability of the tag generated at that generalization level. In order to assess the generated structure and demonstrate the ability of the HACM model to identify abstraction levels in the hierarchy, we have visualized the probabilisty distribution involving the tag "programming" and "xml" in Fig.14 and Fig.15 respectively. We can see that the "programming" tag is mostly used as a very general term. Hence the root

node contains the majority of its probability mass. The tag is also used with "microsoft" and "regex" for its narrower sense of MS programming and programming with regular expressions. This is reflected in the lower-left corner of the Fig.14. On the contrary, the tag "xml" is used mostly as a very specific sense as in "xml programming". It thus appears large at the bottom of the hierarchy. It, however, also used in a more general sense as a data format as in discussion with "atom" and "feed" in height 2. These examples are only spotlights, but they showes that there indeed **are** hierarchical relations among the emerged concepts and it is possible to discover them using more refined probabilistic models. The discovered hierarchy can be used as a basis for further manual refinement for a taxonomy.

The advantage of such a generated taxonomy is that it is dynamically generated from free-style bottom-up annotations and it directly reflects the users' vocabularies. The taxonomy thus can be effectively understood and utilized by the community users. This avoids the drawbacks of the top-down approach to semantic annotation in which the ontology is built before its actual use and therefore may have missmatch with the requirements of its applications and may out-of-sync with the resources the ontology intends to cover. Needless to say, the bottom-up annotation removes the high barrier to entry in top-down semantic annotations because the users need not to have sophiscated knowledge about taxonomy or ontology to make the annotation.

5 Related Work

Semantic annotation is a key problem in the Semantic Web area. A lot of work has been done about the topic. Early work like [16,17] mainly uses an ontology engineering tool to build an ontology first and then manually annotate web resources in the tool. In order to help automate the manual process, many techniques has been proposed and evaluated. [22] learns from a small amount of training examples and then automatically tags concept instances on the web. The work has been tested on a very large-scale basis and achieves impressive precision. [20] helps users annotate documents by automatically generate natural language sentences according to the ontology and let users interact with these sentences to incrementally formalize them. Annother interesting approach is proposed by [21] that utilizes the web itself as a disambiguation source. Most annotations can be disambiguated purely by the number of hits returned by web search engines on the web. [24] improves the method using more sophisticated statistical analysis. Given that many web pages nowadays are generated from a backend database, [19] proposes to automatically produce semantic annotations from the database for the web pages. Information extraction techniques are employed by [23] to automatically extract instances of concepts of a given ontology from web pages. However, these work on semantic annotation follows the traditional top-down approach to semantic annotation which assumes that an ontology is built before the annotation process.

Our work of automatic taxonomy generation from folksonomy can be seen as a method for ontology learning [35] which has lot of related work. [36] gives

a comprehensive review of the state-of-the-art ontology learning methods and places them in a framework for comparision. Most ontology learning methods learn ontology from structured data (e.g. database schema), semi-structured data on the web (e.g. HTML, XML and DTDs) and unstructured data (i.e. text). Very few work exploits the social bookmarks for ontology learning. [37] learns ontology from bookmarks, but the bookmarks used are those personal bookmarks stored on personal PCs that are not shared. Our work learns a taxonomy from the shared social bookmarks.

Much work has been done to help users manage their bookmarks on the (semantic) web such as [27]. [28] gives a good review of the social bookmarks tools available. These tools help make the social bookmarking easy to use but lacks capabilities to derive emergent semantics from the social bookmarks.

Work on emergent semantics [25,26] has appeared recently, for example [38,39,40]. [39] proposes an emergent semantics framework for large scale distributed systems and gives a good example of the framework. It shows how the spreading of simple ontology mappings among adjacent peers can be utilized to incremently achieve a global consensus of the ontology mapping. [40] described how to incrementally obtain a unified data schema from the users of a large collection of heterogeneous data sources. [38] is more related to our work. It proposes that the semantics of a web page should not and can not be decided alone by the author. The semantics of a web page is also determined by how the users use the web page. This idea is similar to our thought. In our work, a URL's semantics is determined from the users' tags. However, our method of achieving emergent semantics is different from [38]. We use a probabilistic generative model to analyze user tags while [38] uses common sub-paths of users' web navigation path.

6 Conclusion and Future Work

Traditional top-down approach to semantic annotation in the Semantic Web area has a high barrier to entry and is difficult to scale up. In this paper, we propose a bottom-up approach to semantic annotation of the web resources by exploiting the now popular social bookmarking efforts on the web. The informal social tags and categories in these social bookmarks is coined a name "folksonomy". We quantitatively studied a data set of the Delicious folksonomy and found that power law distributions exist in the data set. This serves as one possible evidence of the implicit social interactions embeded in the folksonomies. Using a probabilistic generative model to interpret the data set, we derived emergent semantics from the folksonomy data. The semantics of URLs and tags can be represented using discrete probability distributions on derived concepts. The ambiguity of the semantics can be quantitatively measured using entropy values of the distributions. Finally, we show that there indeed are hierarchical relations among the emergent concepts and it is plausible to further identify them if we use more refined probabilistic models. In summary, compared to the top-down approach, the bottom-up approach does not depend on a pre-defined semantic

model to assign semantics to resources but rather derives them from the real usage data. This entitles the approach several advantages such as the low barrier to entry and the tight connection to user vacabularies.

As our work done in this paper is mainly quantitative, future work needs be done more theoretically. We have several topics in our mind that need further exploration. The first one is how the top-down approach and the bottom-up approach may be combined together to leverage both advantages to solve the challenging problem of semantic annotation. This requires innovative thinking and deep insights. An accompanying question about what is the relationship between the representations of semantics in the bottom-up approach and the formal representations in the top-down approach and how we may link them together is also intriguring. Comparing the bottom-up approach in this paper with other probabilistic methods such as LSI [41] and conducting a formal rigorous evaluation is another future topic. Finally, automatically obtaining concept hierarchies from folksonomies is an open, difficult and challenging problem that worth a great effort to attack.

References

1. Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web. Scientific American **284** (2001) 34–43
2. Manola, F., Miller, E.: RDF Primer. W3C Recommendation (2004)
3. McGuinness, D.L., van Harmelen, F.: OWL Web ontology language overview. W3C Recommendation (2004)
4. H.Gennari, J., A.Musen, M., W.Fergerson, R., E.Grosso, W., Crubézy, M., Eriksson, H., F.Noy, N., W.Tu, S.: The evolution of Protégé: An environment for knowledge-based systems development. Technical Report SMI-2002-0943, Stanford Medical Informatics (2002)
5. Bechhofer, S., Horrocks, I., Goble, C., Stevens, R.: OilEd: a reason-able ontology editor for the semantic web. In: Proceedings of the Joint German/Austrian Conference on AI. LNCS 2174 (2001) 396–408
6. Corcho, O., López, M.F., Pérez, A.G., Vicente, O.: WebODE: An integrated workbench for ontology representation, reasoning, and exchange. In: Proceedings of EKAW 2002. LNCS 2473 (2002) 138–153
7. Zhang, L., Yu, Y., Lu, J., Lin, C., Tu, K., Guo, M., Zhang, Z., Xie, G., Su, Z., Pan, Y.: ORIENT: Integrate ontology engineering into industry tooling environment. In: Proc. of the 3rd Intl. Semantic Web Conference (ISWC2004). (2004)
8. Kalyanpur, A., Sirin, E., Parsia, B., Hendler, J.: Hypermedia inspired ontology engineering environment: SWOOP. In: Proc. of the 3rd Intl. Semantic Web Conference (ISWC2004). (2004)
9. Heflin, J., Hendler, J.: Dynamic ontologies on the web. In: Proceedings of the Seventeenth National Conference on Artificial Intelligence (AAAI-2000), Menlo Park, CA, USA, AAAI/MIT Press (2000) 443–449
10. F.Noy, N., Klein, M.: Ontology evolution: Not the same as schema evolution. Knowledge and Information Systems **5** (2003)
11. Kiryakov, A., Ognyanov, D.: Tracking changes in RDF(S) repositories. In: Proceedings of the EKAW 2002, Siguenza, Spain, Springer (2002) 373–378

12. Noy, N.F., Kunnatur, S., Klein, M., Musen, M.A.: Tracking changes during ontology evolution. In: Proc. of the 3rd Intl. Semantic Web Conference (ISWC2004). (2004)
13. Klein, M., Fensel, D.: Ontology versioning for the semantic web. In: Proceedings of the 1st International Semantic Web Working Symposium (SWWS'01), Stanford University (2001) 75–91
14. Klein, M., Fensel, D., Kiryakov, A., Ognyanov, D.: Ontology versioning and change detection on the web. In: Proceedings of the EKAW 2002, Siguenza, Spain, Springer (2002) 197–212
15. Stojanovic, L., Maedche, A., Motik, B., Stojanovic, N.: User-driven ontology evolution management. In: Proceedings of the EKAW 2002, Siguenza, Spain, Springer (2002) 285–300
16. N.F.Noy, M.Sintek, S.Decker, M.Crubezy, R.W.Fergerson, M.A.Musen: Creating semantic web contents with Protege-2000. IEEE Intelligent Systems **2** (2001) 60–71
17. S.Handschuh, S.Staab: Authoring and annotation of web pages in CREAM. In: Proc. of the 11th Intl. World Wide Web Conference (WWW2002). (2002)
18. Kiryakov, A., Popov, B., Ognyanoff, D., Manov, D., Kirilov, A., Goranov, M.: Semantic annotation, indexing, and retrieval. In: Proc. of the 2nd Intl. Semantic Web Conference (ISWC2003). (2003)
19. Handschuh, S., Staab, S., Volz, R.: On deep annotation. In: Proc. of the 12th Intl. World Wide Web Conference (WWW2003). (2003) 431–438
20. Blythe, J., Gil, Y.: Incremental formalization of document annotations through ontology-based paraphrasing. In: Proc. of the 13th conference on World Wide Web (WWW2004), ACM Press (2004) 455–461
21. Cimiano, P., Handschuh, S., Staab, S.: Towards the self-annotating web. In: Proc. of the 13th Intl. World Wide Web Conference (WWW2004). (2004)
22. Dill, S., Eiron, N., Gibson, D., Gruhl, D., R.Guha, Jhingran, A., Kanungo, T., Rajagopalan, S., Tomkins, A., A.Tomlin, J., Y.Zien, J.: SemTag and Seeker: Bootstrapping the semantic web via automated semantic annotation. In: Proc. of the 12th Intl. World Wide Web Conference (WWW2003). (2003) 178–186
23. Etzioni, O., Cafarella, M., Downey, D., Kok, S., Popescu, A.M., Shaked, T., Soderland, S., S.Weld, D., Yates, A.: Web-scale information extraction in KnowItAll (preliminary results). In: Proc. of the 13th Intl. World Wide Web Conf.(WWW2004). (2004)
24. Cimiano, P., Ladwig, G., Staab, S.: Gimme the context: Context-driven automatic semantic annotation with C-PANKOW. In: Proc. of the 14th Intl. World Wide Web Conference (WWW2005). (2005)
25. Maedche, A.: Emergent semantics for ontologies. IEEE Intelligent Systems **17** (2002)
26. Aberer, K., et.al: Emergent semantics principles and issues. In: Proc. of Database Systems for Advanced Applications. LNCS 2973 (2004)
27. Kahan, J., Koivunen, M.R., Prud'Hommeaux, E., Swick, R.R.: Annotea: An open RDF infrastructure for shared web annotations. In: Proc. of the 10th Intl. World Wide Web Conference. (2001)
28. Hammond, T., Hannay, T., Lund, B., Scott, J.: Social bookmarking tools (i) - a general review. D-Lib Magazine **11** (2005)
29. Mathes, A.: Folksonomies - cooperative classification and communication through shared metadata. Computer Mediated Communication, LIS590CMC (Doctoral Seminar), Graduate School of Library and Information Science, University of Illinois Urbana-Champaign (2004)

30. Udell, J.: Collaborative knowledge gardening. InfoWorld, August 20 (2004)
31. Merholz, P.: Metadata for the masses. http://www.adaptivepath.com/publications/essays/archives/000361.php, accessed at May, 2005. (2004)
32. Adamic, L.A., Huberman, B.A.: The web's hidden order. Communications of the ACM **44** (2001)
33. Hofmann, T., Puzicha, J.: Statistical models for co-occurrence data. Technical report, A.I.Memo 1635, MIT (1998)
34. G.A.Miller: WordNet: A lexical database for english. Communications of the ACM **2** (1995)
35. A.Maedche, S.Staab: Ontology learning for the semantic web. IEEE Intelligent Systems **16** (2001)
36. M.Shamsfard M, A.: The state of the art in ontology learning: a framework for comparison. Knowledge Engineering Review **18** (2003)
37. J.J.Jung, Y.H.Yu, S.S.Jo: Collaborative web browsing based on ontology learning from bookmarks. In: Proc. of the Intl. Conference of Computational Science (ICCS2004). (2004)
38. W.I.Grosky, D.V.Sreenath, F.Fotouhi: Emergent semantics and the multimedia semantic web. SIGMOD Record **31** (2002)
39. Aberer, K., Cudre-Mauroux, P., Hauswirth, M.: The chatty web: Emergent semantics through gossiping. In: Proc. of 12th Intl. Conf. on World Wide Web (WWW2003). (2003)
40. Howe, B., Tanna, K., Turner, P., Maier, D.: Emergent semantics: Towards self-organizing scientific metadata. In: Proc. of the 1st Intl. IFIP Conference on Semantics of a Networked World: Semantics for Grid Databases (ICSNW 2004). LNCS 3226 (2004)
41. W.Furnas, G., Deerwester, S., T.Dumais, S., K.Landauer, T., A.Harshman, R., A.Streeter, L., E.Lochbaum, K.: Information retrieval using a singular value decomposition model of latent semantic structure. In: Proc. of the ACM SIGIR'88, Grenoble, France (1988) 465–480